VŠB - Technická univerzita Ostrava

Fakulta elektrotechniky a informatiky

Katedra informatiky

Analýza vývoje komunit v rozsáhlých sociálních sítích

Analysis of Community Evolution in Large Social Networks

2015 Bc. Karel Gazárek

VŠB - Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky
Katedra informatiky

Zadání diplomové práce

Student: **Bc. Karel Gazárek**

Studijní program: N2647 Informační a komunikační technologie

Studijní obor: 2612T025 Informatika a výpočetní technika

Téma: Analýza vývoje komunit v rozsáhlých sociálních sítích
Analysis of Community Evolution in Large Social Networks

Zásady pro vypracování:

Cílem práce je provedení průzkumu existujících přístupů, návrh a implementace vybrané nebo vlastní metody a aplikačního prostředí pro experimenty.

1. Průzkum a popis existujících přístupů.
2. Návrh a implementace vybrané nebo vlastní metody.
3. Návrh a implementace počítačové aplikace pro provádění experimentů.
4. Návrh, realizace a hodnocení experimentů.

Seznam doporučené odborné literatury:

[1] Mark Newman. Networks: An Introduction. Oxford University Press 2010
[2] Dále podle pokynů vedoucího diplomové práce.

Formální náležitosti a rozsah diplomové práce stanoví pokyny pro vypracování zveřejněné na webových stránkách fakulty.

Vedoucí diplomové práce: **Mgr. Miloš Kudělka, Ph.D.**

Datum zadání: 16.11.2012
Datum odevzdání: 07.05.2015

doc. Dr. Ing. Eduard Sojka
vedoucí katedry

prof. RNDr. Václav Snášel, CSc.
děkan fakulty

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

V Ostravě dne 30. července 2015

.....

podpis

Poděkování

Na tomto místě bych rád poděkoval doc. Mgr. Miloši Kudělkovi, Ph.D. za metodické vedení této práce, za jeho trpělivost a dobré rady. Dále bych chtěl poděkovat mé partnerce, rodině a přátelům za podporu během studia.

Abstrakt

Téma diplomové práce se zaměřuje na problematiku Analýzy vývoje komunit v rozsáhlých sociálních sítích. Práce je rozdělena do několika kapitol zahrnující jak vymezení teoretické, tak následně praktickou realizaci vybraných experimentů a statistik na základě kapitol teoretické části práce. Stejně jsou především kapitoly týkající se vymezení základních řešení pojmů, metod a analýz včetně jejich popisu, na toto následně navazuje praktická část s uvedením zdroje dat pro zpracování experimentů a statistik, jejich vyhodnocení a následně formulace závěrů a zjištěných skutečností.

Klíčová slova

Sociální sítě, komunity, internet, analýza sociální sítě

Abstract

The theme of the thesis focuses on analysis of the development of communities in large social networks. The work is divided into several chapters on how inclusive the definition of theoretical and then the practical realization of selected experiments and statistics based on theoretical chapter. As well as particularly the chapter on definitions of basic concepts and solving methods and analyzes, including their description, and that this then leads on to the practical part whit the source data for the processing of experiments and statistics, assessment and subsequent formulation of conclusions and findings.

Key words

Social networking, community, internet, social network analysis

Obsah

1. Úvod

Téma diplomové práce se zaměřuje na problematiku Analýzy vývoje komunit v rozsáhlých sociálních sítích. Současné webové sociální sítě jsou v současné době stále populárnější. Stávající se prostorem, v jehož rámci lidé tráví velké množství času, je možné konstatovat, že pro některé jsou sociální sítě druhým životem. Je to také prostor, kde lidé nejenom komunikují, ale také navazují přátelství, zakládají specifické komunity a zájmové skupiny. V rámci sociální sítí, je možné získat o jejich uživatelích množství informací, proto je například využívají i personalisté a další zainteresované osoby, jsou také zdrojem marketingových informací a dalších dat, která vypovídají o jednotlivých uživatelích.

Lidé se obecně zajímají o nové informace, k tomuto je možné použít metody dobývání dat. Je uplatňována také analýza sociálních sítí, která je schopna získat informace ze struktury sociálních sítí. Povaha dat ze sociálních sítí umožňuje získávat data z webových aplikací a umožňuje tak hledání nových metod, které by mohly efektivněji vyhledávat a získávat potřebná data a informace. V rámci cílů diplomové práce je nalezení vhodné metody pro získání a dobývání znalostí ze sociální sítě a následně je ověřit na získaných datech.

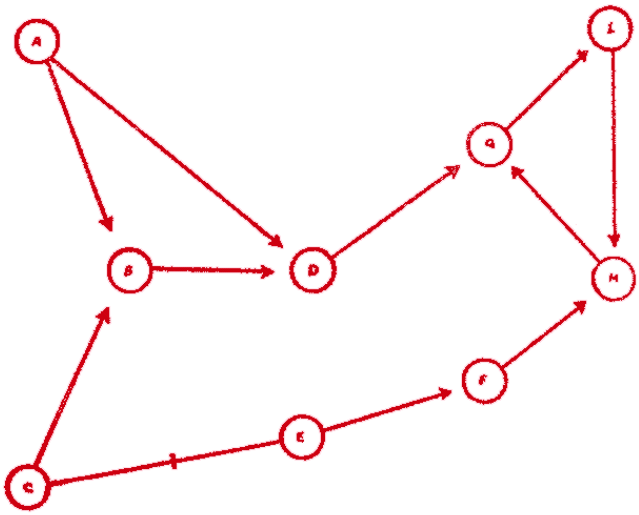
Tato práce je rozdělena do dvou velkých celků, a to na teoretickou část práce a praktickou část práce. Kdy v teoretické části se zabývám teoretickými koncepty a vymezením základních problémů, které jsou následně rozpracovány v praktické části práce. Následné kapitoly v praktické části práce popisují a analyzují metody získání dat a analýzu vývoje komunit na sociálních sítích. Sociální síť je tak ústředním tématem práce, má svoji vlastní terminologii a je doplněna o některé pojmy z grafové teorie.

Závěrem práce jsou prezentovány a okomentovány výsledky všech experimentů provedených aplikací, která je nedílnou součástí pro implementovaný algoritmus.

2. Teoretická část práce

2.1 Pojem sociální síť

Sociální sítě se v současném globalizujícím světě staly fenoménem doby. Vznik a rozvoj sociálních sítí v posledních letech změnily nejenom lidské chování, ale i myšlení v mnoha směrech díky rychlému přenosu informací a to především u mladších generací. Když se řekne „sociální síť" velmi mnoho lidí si v dnešní době představí webové služby, jako jsou například Facebook, Tweeter, MySpace nebo největší profesní síť na světě LinkedIn, těmto sociálním sítím se podrobněji věnuji v další části práce. Ale nebylo tomu vždycky tak. V druhé polovině dvacátého století tento pojem ve vědeckém kontextu poprvé užil profesor ekonomické university John Arundel Barnes v roce 1954. Profesor Barnes zkoumal sociální vazby mezi rybáři v Norské vesnici Bremnes na ostrově Bomlo. Tyto vazby zkoumal téměř dva roky a jeho závěrem byla myšlenka, že celou společnost můžeme definovat jako množinu bodů, z nichž některé jsou propojeny linkami. Tato množina linek a bodů pak utváří celkovou síť vztahů tj. sociální síť. Tato myšlenka, ale vznikla ještě dříve a to ve 30. letech dvacátého století. Zabýval se jí americký sociolog doktor psychiatr Jacob Levi Moreno, který roku 1934 představil první grafickou interpretaci těchto vazeb a nazval ji sociogram. [8]



Obrázek 1: Sociogram Moreno

(zdroj: http://www.cmu.edu/joss/content/articles/volume1/Freeman.html)

2.2 Historické kořeny sociálních sítí

Historie vzniku webových sociálních sítí se váže k počátkům vzniku předchůdce dnešního internetu ARPA NETU, který vzniknul roku 1969. V roce 1971 Ray Tomlinson odeslal první email. Vznik elektronické pošty lze považovat za počátek sociálních vazeb ve virtuálním prostoru. Avšak za počátek sociálních sítí lze považovat, až rok 1978 v tomto roce byl poprvé spuštěn systém BBS, který byl od roku 1972 provozován jako experiment, který zkoumal, jak se lidé budou chovat při vyměňováni informací prostřednictvím počítače. BBS byl soubor elektronických nástěnek, prostřednictvím něho si mohli uživatelé poprvé vyměňovat informace mezi sebou. Tento systém však nebylo velice efektivní, z důvodu, že v určitou chvíli mohl být připojen jen jeden uživatel a mimo to byl tento systém velice pomalý. [15]

Dalším prvkem pro komunikaci v reálném čase například prostřednictvím „chatu“, jehož první možnou variantou byl IRC – Internet Realy Chat, jež v roce 1988 spustili dva finští programátoři, tento program v této době patřil pouze Finsku. V současné době je IRC největší existující síti převážně pro Evropu. Mezi některé významné české IRCnet servery patří například **irc.felk.cvut.cz** (IP adresa 147.32.80.79, port 6667), **irc.ipv6.cesnet.cz** (IPv6 komunikace, port 6667). [16]

Dalším, důležitým krokem který vedl ke vzniku sociálních sítí, bylo zveřejnění prvních internetových stránek. To umožnil anglický vědec Tim Berners-Lee, který v roce 1989 začal pracovat na globálním hypertextovém projektu známém jako World Wide Web. Ten byl vytvořen pro podporu spolupráce lidí, v něm jim umožňoval kombinovat jejich vědomosti v rámci sítě hypertextových odkazů. Po naprogramování webového serveru a webového prohlížeče tak v roce 1991 publikoval první internetové stránky již zmíněný vědec Tim Brenes-Lee a tím pádem udělal další krok, který vedl ke vzniku sociálních sítí, jak je známe dnes. [17]

V roce 1995 vznikla první sociální síť classmates.com, kterou vybudoval Randy Conrad. Díky těmto webovým stránkám bylo možné uživatelům hledat, registrovat a udržovat vazby mezi spolužáky studenty a jinými lidmi. V současné době má web zhruba 40 milionu aktivních uživatelů, z nichž většina je ze Spojených státu amerických a Kanady. Tento Web se stal velmi úspěšným a to díky myšlence tvorby jednotlivých vztahů mezi uživateli, díky tomuto se stal vzorem pro dnes nejznámější internetové sítě jako je například MySpace, LindekIn a Facebook, kterým se budu podrobněji věnovat v dalších kapitolách.

V roce 1997 byl poprvé použit termín „weblog“, který přestavil Jorn Barger. Tento termín označoval seznam odkazů o politickém, kulturním a technologickém dění, které byly dle něj zajímavé a o které se chtěl s veřejnosti podělit. Z toho se vyvinula služba, se kterou se asi nejvíce proslavil a největších úspěchů dosáhl Blogger.com. Ten vznikl roku 1999 a dodnes je největším poskytovatelem jednoduchého publikování delších textů. [18]

Mezi léty 1997 a 2001 fungoval projekt zvaný SixDegrees.com, který sledoval myšlenku sítě kontaktů. Pojmenován byl po teorii šesti stupňů odloučení, která předpokládá, že každé dvě osoby na zemi jsou spojeny řetězcem šesti navzájem si známých lidí. SixDegrees nabízela možnost zasílat zprávy a objekty na nástěnku lidem z prvního až třetího stupně odloučení. [24]

Obrázek 2: Logo SixDegrees

(Zdroj: http://blog.afridesign.com/wpcontent/uploads/2010/09/sixdegrees_logo.jpg)

2.3 Sociální sítě

Sociální sítě jsou sítě, ve kterých jsou vrcholy reprezentovány osobami nebo skupinami lidí a hrany jsou reprezentovány formou sociální interakce mezi nimi. Jako je například přátelství. V informatice jsou vrcholy považované za uzly (nodes) a hrany za spojení (connections). V jiných oborech se vžily jiné názvy. Například sociologové používají pro označení vrcholů aktéry (actors) a pro označení hran vazby (ties).

Jde o virtuální propojení skupiny lidí, díky niž je možno sdílet různé typy informací, jako například odkazy, obrázky, videa, fotky atd., které se vzájemně ovlivňují. Celosvětově existuje několik desítek sociálních sítí, jejichž základem je sdílení informací na internetu mezi přáteli jinými slovy řečeno jejími uživateli. Velice často se nesprávně zaměňují pojmy „sociální síť“ a „komunita“, kdy sociální síť značí zcela volnou, náhodnou interakci, vzniklou na základě času a prostoru, oproti tomu komunita znamená sdílení informací s podobně zaměřenými lidmi a podobnými zájmy, kdy jako příklad je možné uvést FACEBOOK.

V současné době existují dvě základní rozdělení sociálních sítí:

- Osobní – ty jsou určeny pro sdílení osobních informací a obsahů, je zaměřen na specifický obsah nebo oblast (fotky, hudba, videa)
- Profesní – cílem profesní sítě, je získání konkrétních informací (některé hledají konkrétní informace, jiné spravují kontakty) [7]

2.4 Webové sociální sítě a jejich vymezení

Webové sociální sítě představují jeden z hlavních pojmů v nové fázi internetu, jež je označována jako Web 2.0. Je to systém, založený na postupné evoluci, není tedy systémem vzniklým náhle, ale jde o určitý vývojový proces. Sociální sítě jsou založené na koncepci Web 2.0, byly poprvé prezentovány Timem O´Reillym v roce 2004, kde uživatelé vytváří vlastní obsah dané stránky, komunikují a sdílejí informace, tím vytváří hodnotu webu. Základním předpokladem aplikací vyhovujících standardu Web 2.0 je modulace, zajišťující implementaci nových funkcí, trendů a technologií na základě

požadavků uživatelů. Ty jsou zpravidla do aplikací sociálních sítí integrovány ihned po svém uvedení a úspěšném otestování jejich funkčnosti. [26]

Mezi prvky, které jsou typické pro Web 2.0., a pro nové webové sociální sítě se uplatňuje například jev, kdy se tento web stále více stává platformou, kde se začínají prosazovat stále více internetové aplikace v souvislosti například s obchodem, marketingem, právě oblast marketingu mě zaujala natolik, že se jí budu v další kapitole věnovat podrobněji.

Dalším aspektem je začlenění kolektivní inteligence do webových sociálních aplikací, kdy samotní uživatelé často tvoří většinu obsahu webové sociální aplikace, které se vyznačují následujícími charakteristikami:

- tvorba velké části obsahu uživateli
- kategorizace obsahu
- používání značek v obsahu
- provázání obsahu odkazy [1]

2.5 Moderní sociální sítě

V této podkapitole se věnuji moderním sociálním sítím současné doby, typickým znakem moderních sociálních sítí je komunikace, která probíhá prostřednictvím přímé nebo hromadně komunikace formou odeslané zprávy. Dalším aspektem sociálních sítí je vytvoření profilu, kde si uživatelé mohou nahrát svou fotku a další informace jako například osobní údaje, záliby, oblíbené koníčky atd. dále mohou uzavírat přátelství a komunikovat s různými uživateli taktéž připojenými k síti. Sociální síť je tedy založena na propojování jednotlivých uživatelů s různými vazbami, ať už se jedná o práci, školu, prestiž atd.

Mezi tyto sítě patří například již nejznámější síť a to:
Facebook – vznikl v roce 2004, jeho zakladatelem je vysokoškolský student Hardvardské univerzity Mark Zuckerberg spolu s Dustinem Moskovitzem, Chrisem Hughesem a Eduardem Saverinem. V prvopočátku fungoval Facebook jako program, který propojoval komunikaci mezi vysokoškoláky Harvardu, ten sklidil obrovský úspěch, během jednoho měsíce od spuštění se

na tuto sociální sít přihlásilo 19 500 studentů s původním názvem thefacebook.com. [9]

V současné době má Facebook více než 1,2 miliardu aktivních účtů z toho v České Republice je zhruba 4,2 milionů uživatelů. Typickým znakem Facebooku je usnadnění komunikace a zprostředkování kontaktu mezi lidmi. Každý uživatel, který chce mít svůj profil na Facebooku si musí založit účet prostřednictvím registračního formuláře, který je zdarma k dispozici na interentu. [10]

LinkedIn – je sociální sít zaměřená na profesní kontakty, jejimž hlavním cílem je „professional networking“, vpřekladu sdružování profesionálů. Tato sociální síť má více než 40 milionu členů. LinkedIn byl spuštěn v roce 2003. V dnešní době má tato síť 120 milionu registrovaných uživatelů ve více jak 196 zemí světa. Služba LindekIn je orientována na profesionály ve svém oboru nebo taktéž podnikatelé, kteří si díky této síti mohou vyměňovat zkušenosti, vědomosti, informace a předávat si rady a doporučení v rámci svých pracovních zájmů. [11]

Obrázek 3: Logo LinkedIn

(Zdroj: http://www.lupa.cz/clanky/linkedin-pro-zacatecniky/)

Twitter – byl spuštěn v roce 2006, v současné době patří mezi nepoužívanější mikrology, na celém světě, jeho služeb využívá více jak 140 milionu uživatelů. Posláním twittru je přenést informace v reálném čase. Díky své jednoduchosti se stal velice populárním, všichni registrovaní uživatelé mají možnost odesílat nebo sledovat krátké zprávy takzvané tweetů, které nesmějí obsahovat více jak 140 znaků. [12]

Obrázek 4: Logo Twitter

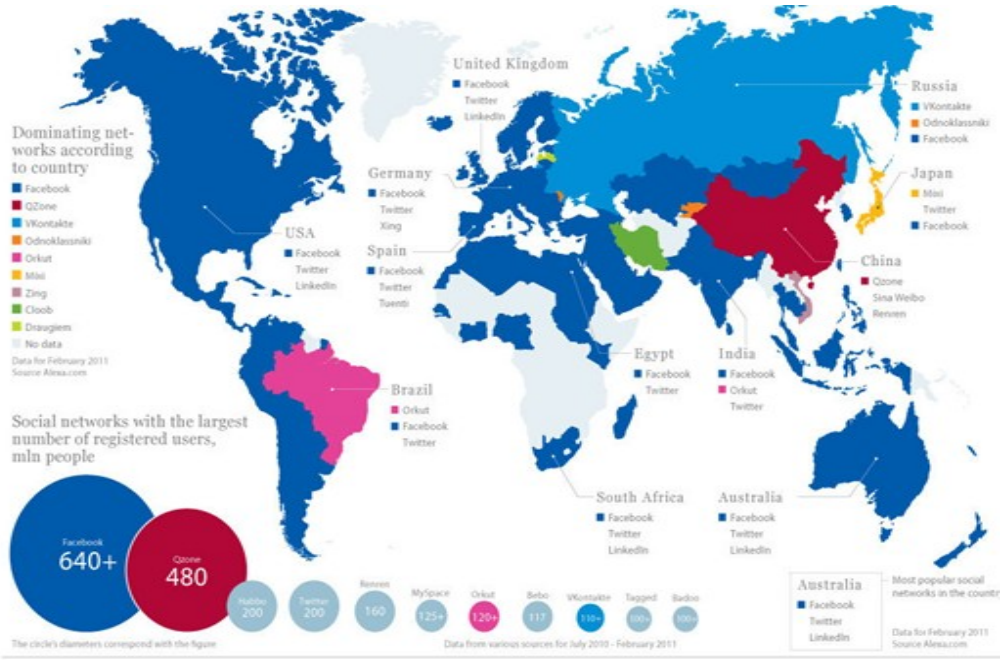
(Zdroj: http://www.business2community.com/twitter/use-twitter-business-2-01077755)

MySpace – vznikl v roce 2003 v Kalifornii, jednalo se o první sociální síť podporující internetový marketing. Zajímavostí je, že tato síť byla vytvořena za neuvěřitelných 10 dní, MySpace je napodobeninou Friendster.com, která vznikla v roce 2003, je to první síť, díky které došlo k propojení reálných přátel na internetu. MySpace je sociální síť známá po celém světě, je zaměřena na poslech hudby. [13]

Obrázek 5: Logo MySpace

(Zdroj: http://en.wikipedia.org/wiki/Myspace)

Služby těchto vybraných sociálních sítí jsou využívány po celém světě, k lepší orientaci jsem si dovolil přiložit tabulku mapy světa, ve které jsou zobrazeny jednotlivé země využívajících služeb sociálních sítí. Z tohoto obrázku je patrné, že využívání sítí se mění stát od státu například v Japonsku je nejoblíbenější sítí Mixi v Rusku pak VKontakte, ale mezi nejoblíbenější a nepoužívanější sítí na světě je Facebook za ním se řadí Twitter a v neposlední řadě LindekIn. [14]



Obrázek 6: Mapa světových sociálních sítí

(Zdroj: http://www.justit.cz/wordpress/2013/05/29/infografika-svetova-mapa-socialnich-siti-z-roku-2011/)

Výše popsané sociální sítě jsou známé především jako zahraniční sociální sítě celosvětového zájmu, ale i čeští uživatelé mají možnost využívat služeb českých sociálních webů. Mezi nejznámější české sociální sítě patří:

- **Líbímseti.cz** – jedná se o zábavný server, který byl založen v roce 2002.
- **Spolužáci.cz** – je komunitní server, na němž se nachází databáze tříd různých základních, středních škol, gymnazií či učilišť v České republice. [22]
- **Lidé.cz** - tato síť patří mezi největší a nejoblíbenější sítí v ČR s více než 140 tisíci uživateli denně. Na této stránce se může uživatel seznámit, s jinými uživateli, popovídat si či si založit svůj vlastní blog.

Je však nutno podotknout, že s příchodem Facebooku klesl počet uživatelů, výše popsaných serverů. Tento propad je evidován od roku 2009, kdy Facebook vstoupil na český trh, důvodem oblíbenosti tohoto serveru je nejspíš fakt, že všechny funkce výše zmíněných komunikačních portálů provozuje.

3. Sociální sítě a marketing

V této práci jsem se snažil podívat na dané téma z marketingového pohledu, hlavně z důvodu, že marketing v kombinaci se sociálními sítěmi hraje v současné době velkou roli, i když si to například řada z nás v prvních chvílích neuvědomuje. V následujících podkapitolách se budu snažit vysvětlit tento vztah.

V dnešním moderním světě hraje důležitou roli pojem E-commerce v souvislosti se sociálními sítěmi vymezuje zejména oblasti od distribuce, nákupu, prodeje, marketingu a servírování produktů a služeb prostřednictvím elektronických systémů. Elektronická komerce dnes zahrnuje nejenom nákup a prodej po internetu, ale také například online platby, e-marketing, elektronické výměny dat (EDI), automatické sběry dat a mnoho dalších možností. E-komerce využívá komunikačních technologií, zejména především www stránky, e-shopy, databáze, e-maily, vyhledávače a jiné technologie k nákupu svých produktu. Pojem e-commerce zastřešuje širokou oblast nástrojů a principů. Jeho úkolem je vhodně tyto nástroje kombinovat tak, aby co nejlépe odpovídaly požadavkům firem na jejich komerční působení na internetu. [2]

Veškeré aktivity související s prací s webovými sociálními sítěmi, jež jsou určeny pro koncové zákazníky – cílové skupiny uživatelů, by měly být směřovány k navázání nebo prohloubení důvěry v provozovatele a jeho produktů tak, aby se zvyšovalo jednak povědomí a zájem o dané služby takového webu nebo sociální sítě. Za klíčové jsou následující faktory:

- grafická prezentace, působení na náladu zákazníka,
- kvalitní obsah, kvalitní popis zboží,
- uživatelská přívětivost, tedy to, jak snadno se ovládá daný webová stránka, její aplikace, sociální síť.
- Například lze měřit, jak složitý je samotný proces hodnocení zájmů a aktivit uživatelů,
- zpětná vazba

Komunikační a informační technologie přinášejí nové portfolio možnosti pro využití v řadě odvětví, stejně tak i v oblasti marketingových aktivit podniků. Efektivní online komunikace a profesionální webová prezentace může být dnes konkurenční výhodou řady firem a zanedlouho i nezbytností, kterou budou perspektivní podniky používat. [3]

Marketing neustále prochází změnami, právě proto by si firmy měly uvědomit, zda jejich marketing je efektivní a zároveň znovu promyslet svoji marketingovou a komunikační strategii. Zahrnutí internetu a prezentace v kombinaci se sociálními sítěmi do aktivit firmy vyvolává celou řadu výhod, které můžeme vymezit následujícím způsobem [4]:

- zvýšení počtu komunikačních kanálů
- zvýšení dostupnosti informací
- vytvoření moderního image
- moderní podpora prodeje

Webová prezentace podniku je základním aspektem působnosti v oblasti internetu a elektronických marketingových prostředků. Jde v podstatě o řadu vzájemně provázaných webových stránek, na kterých mohou být publikovány základní informace o firmě, jejím poslání, oboru činnosti, nabídce zboží či služeb a jejich specifikace a srovnání, seznam provozoven, kontakty na odpovědné osoby a podobně. Obecně se jedná o jednoduchý způsob jak zákazníkům sdělit to, co potřebují vědět prostřednictvím internetových portálu.

Webové sídlo firmy tak představuje portál, který vytváří rozhraní pro návštěvníky.

Formy propagace můžou mít různou podobu, mezi nejznámější patří ty, které využívají tradiční poštovní zásilky či elektronické maily. V případě jasně vymezené cílové skupiny je velmi efektivní, zhledem k tomu, že dává najevo přímý zájem. Příjemce dopisů či mailů má tímto aktualizované veškeré informace o uživateli, případně o kandidátovi.

Podstatným znakem v rámci vytváření pozitivní image jsou vizuální znaky. Patří mezi ně různé symboly, loga či použité barvy, www stránky, aktivity prostředním internetových sociálních sítí jako např. FACEBOOK atd.. [5]

3.1 Sociální sítě a jejich využití v oblasti marketingu

Typickým příkladem dnešní doby může být virální marketing, kdy se díky doporučení obsahu nebo konkrétního odkazu, které jsou přeposílány mezi přáteli, v sociální sítí např. FACEBOOKU, propaguje produkt, nebo služba, která není na první pohled vidět. Důležitým jevem je, aby sdělení pobavilo nebo upoutalo pozornost a uživatelé či čtenáři si je tak mezi sebou přeposílali. Při realizaci virálního marketingu se používají techniky, které šokují veřejnost. Mezi největší internetové servery pro sdílení videí (virilního marketingu) patří server YouTube, který momentálně spadá pod korporaci Google. V České republice patří mezi oblíbený nástroj, obzvlášť kvůli rozsáhlému obsahu. [19]

Webové sociální sítě se mohou zaměřovat na prezentaci, jak bylo naznačeno výše, nebo také na budování sociálních vazeb mezi jednotlivými uživateli. Základem jsou dostupné webové stránky s vlastním profilem každého člena sociální sítě a vybudovaných vazeb mezi ostatními uživateli.

Při využití tohoto marketingového nástroje jsou používána všechna dostupná elektronická zařízení. Jedná se především o internet, mobilní telefony, tablety atd.. Online marketing staví na odhadu chování a vnímání cílové skupiny. Na základě těchto faktorů se snaží vytvořit různé strategie, jak se k uživatelům co nejlépe přiblížit, tak aby byla případná propagace co nejefektivnější. Komunikace postavená na základě tohoto marketingového nástroje, je komunikací přímou, díky niž je možné získat okamžitou odezvu od uživatelů.

Internet je hybnou silou moderní doby, lidé zde tráví většinu svého volného času ve srovnání s ostatními médii, je zde jasnější věkové rozhraní uživatelů. Atraktivita internetu se zvyšuje nejen díky narůstajícímu počtu uživatelů, ale také kvůli rozvíjejícímu se internetovému trhu produktů a služeb. Stále více moderních firem využívá internetu právě pro své reklamní kampaně, které

poskytují pro příjemce nejenom zajímavost a dostupnost, ale také i interaktivnost. Velkou výhodou internetového marketingu je množství a dostupnost uživatelů. Umožňuje oslovit mnohem více lidí, než tradiční reklamní média za minimum stanovených nákladů.

Svou úlohu, zde sehrávají i mobilní zařízení ty představují jeden z rychle se rozvíjejících nástrojů marketingové komunikace. Takzvaný mobilní marketing se vyznačuje právě cílením na konkrétní zákazníky. Momentálně je možné pracovat se socio-demografickými údaji, směřující na zájmy lidí, což zajišťuje mnohem efektivnější práci.

Zájem veřejnosti a oblíbenost získávají tzv. QR kódy, které rozšiřují celkovou nabídku, jedná se o speciální graficky zpracovaný kód, který může být u inzerátu v časopise, televizi nebo na vlakových zastávkách. Při namíření fotoaparátem na tento speciální kód, tak uživatele přímo odkáže na webové stránky či nějakou akční nabídku konkrétního subjektu. [2]

4. Analýza vlastností sociálních sítí

Tato kapitola je věnována "Analýze sociálních sítí", se kterou pracuje ve svém odborném textu pan M.E.J. Newman s názvem "The structure and of complex networks".

Analýza sociálních sítí, se zabývá především analýzou sociální struktury, která se sestavuje z různých entit a vazeb určitého typu, které se vzájemně propojují. Hlavním cílem je objevit popsat a analyzovat vzorce ze sociálních sítí.

4.1 Bezškálovitá síť

Bezškálovitá síť, bývá také často označována jako scale-free network, jde o síť, kde je distribuce uzlů stupňována podle mocninného rozdělení, zásadně se tedy odlišuje od rozdělení v náhodně vybraných grafech, kdy mají tyto grafy tvar zvonovité křivky s jasně vymezeným maximem pro průměrnou hodnotu stupně. Oproti tomu z mocninného rozčlenění grafů žádnou specifickou hodnotu z jejího grafu není možné zjistit. U mocnineho rozčlenění stupňů se uplatňuje jednoduché pravidlo, a to takové, že čím větší je stupeň, tím menší je četnost těchto uzlů. [21]

Příklady bezškálovitých sítí:

- sociální sítě
- sítě interakcí proteinů
- počítačové sítě [20]

4.1.1 Shlukování

Mezí zásadní rozdělení, které je patrné mezi sociálními sítěmi v souvislosti s náhodnými grafy je fakt, že představují mnohem vyšší "shlukování" jinými slovy tranzitivitu C, velice často tak může docházet k situacím, kdy za předpokladu, že existuje vazba uzlu A do uzlu B, a také do uzlu C, existuje velká pravděpodobnost, že mezi uzly B a C také vzniklá vazba. Díky tomuto spojení může existovat, mnohem větší výskyt trojúhelníků v grafech, nebo také může docházet k situacím, kdy přítel přítele, je za tohoto předpokladu i můj

přítel, pokud sítě řeší kupříkladu přátelství. Takovouto vlastnost lze nazvat shlukovací metodou pro každý uzel.

Autor M.E.J. Newman se také zabývá ve svém publikovaném odborném textu, pojmem "Shlukovací koeficient" což je číslo od 0 do 1, které značí pravděpodobnost, kdy dva uzly, které jsou sousedy uzlu I, budou s největší pravděpodobností sousední. Pokud chceme vypočítat hodnoty v rámci celé sítě, je možné použít klasický aritmetický průměr shlukovacích koeficientů u všeobecných uzlů dané sítě. [21]

4.1.2 Organizace propojení uzlů

Propojení uzlů znázorňuje spojení mezi uzly stejných typů, ne však typů rozdílných. Pod typem uzlu je možné si představit například typy lidí, nebo národnosti uživatelů. V sociálních sítích se vyskytuje převážně uspořádané propojení.

Uspořádanost propojení je možné odvodit od matice, ve které sloupce a řádky tvoří samostatné typy a prvky, které znázorňují rozměry vazeb mezi uzly daných typů v návaznosti na jejich celkový počet vazeb. V rámci kvantifikace, pro diskrétní atributy je nejčastěji používaným koeficientem "koeficient uspořádanosti", kde *Tr* vyjadřuje součet prvků jak na diagonále, tak celkový součet prvků u matice. Výsledný koeficient může být také nulový, a to v případě, kdy dochází k náhodným propojením, oproti tomuto kladný výsledek nastává v případě, že dojde k uspořádanému propojení. Pokud dojde k plnému propojení v rámci totožných typů je = 1. [21]

4.1.3 Korelace stupňů

Zvláštním typem korelace stupňů, může být sledování množství propojení v závislosti mezi stupni uzlů. To může nastat za předpokladu, že se uzly s větším množstvím vazeb častěji spojují s podobnými uzly, nebo za situace, kdy se tyto uzly spojují s menšími vazbami. Tento fakt je možné znázornit pomocí výpočtu "Paersonova korelačního koeficientu", který je určen pro hodnoty stupňů z určitých dvojic, mezi kterými existují vazby. "Paersonův korelační koeficient", nabývá jak kladných tak záporných hodnot. Koeficient je kladný v případě, že v síti existuje preference spojení dle stupně, to se stává v mnoha případech u sociálních sítí. Pokud se zaměříme na jiné sítě než sociální, koeficient většinou nabývá záporných hodnot. [21]

4.2 Kohezivní skupiny a jejich vymezení

Pokud se podíváme na strukturu sociálních sítí, zjistíme, že jejich hlavní charakteristikou je fakt, že nejsou homogenní neboli stejnorodé. Kohezní skupiny, nebo jinými slovy také komunity jsou složeny z aktérů, mezi kterými jsou relativně silné, přímé, často také pozitivní vazby.

V dostupné literatuře najdeme zcela jistě několik přístupů k definici kohezivních skupin avšak "oponenta" je tím zcela nejjednodušším. Jedná se o maximálně spojitý podgraf, ve kterém jsou všichni členové vzájemně spojení nějakou z cest, díky ní může probíhat vzájemné spojení. Každá komponenta má určité vlastnosti, jde hlavně o počet členů a poloměr. Při vyhledávání různých komponent je možné, že je k dispozici ihned několik komponent, u kterých existuje ve většině případů jedna hlavní komponenta a určité množství izolovaných aktérů. [21]

4.3 Vizuální analýza sociálních sítí

Vizuální analýza sociálních sítí je založena na hledání zajímavých struktur anebo různých typů vztahů ve vizuální prezentaci sociální sítě. Nejpoužívanějším nástrojem pro vizualizaci je graf. Pokud chceme, aby byl graf co možná nejpřehlednější, je možné použít speciální algoritmy, které se zabývají nalezením rozestavěných uzlů. V grafu je možné zachytit různé

vlastnost od tloušťky hran, která může představovat váhu vazeb, až po různé popisky, barvy nebo velikosti podle některého z atributů.

4.4 Další analytické metody zpracování rozsáhlých dat na sociálních sítích

Využití GPU pro analýzu dat a jiná zpracování časově velmi náročných úkolů je v současném modernizujícím se světě nastupujícím trendem, neboť umožňuje provádět dříve nepředstavitelné úlohy na běžných počítačích a v reálném čase, a v zapojování těchto procesů do největších počítačů světa. Hlavním bodem je tedy návrh a implementace metod redukce dimenze, shlukové analýzy, hledání vzorů, analýzy obrazu a dalších na GPU a jejich využití při analýze dat z různých oblastí jako je například: analýza sociálních sítí a komunit a predikce jejich chování, detekce zájmových oblastí v obrazových datech, podobnost obrázků a dokumentů, a mnoho dalších.

4.5 Analýza sociálních sítí

V současnosti jde o oblast s velkým a velice zajímavým výzkumným potenciálem. Současné technologie umožňují ukládání velkého množství informací spojeného s přímou nebo nepřímou interakcí mezi lidmi. Mezi velmi intenzivně zkoumané problémy v této oblasti patří především dynamika související s vývojem sítě v čase. Jde např. o analýzu šíření informací v síti, detekci komunit a jejich vývoje, identifikace chování a rolí v síti, sledování vývoje obsahu spojeného s interakcemi v síti.

Dobrým vzorem pro aktuálně zkoumané sociální sítě může být veřejná databáze DBLP, která poskytuje byť neúplné, ale vysoce relevantní informace o publikačních aktivitách v oblasti Computer Science, tento systém je používán po celém světě. Mezi nedostatky toho systému je fakt, že poskytuje pouze informace z blízkého okolí jednotlivých autorů. To je zapříčiněno hlavně tím, že rozsah sítě nedovoluje jednoduše aplikovat běžné výpočetní metody tak, aby poskytovaly výsledky v požadovaném čase. [23]

4.6 Centrality

Centralita je základní metrika pro analýzu vrcholů. Je to hodnota, která popisuje jak je vrchol v síti významný. Nelze však přesně definovat jak je vrchol v síti významný. Proto každý typ centrality může určit významnost vrcholu jinak.

Centrality lze rozdělit na tři základní druhy a to na Stupňovou centralitu (Degree Centrality), Centralitu blízkosti (Closeness Centrality) a Centralitu mezilehlosti (Betweenness Centrality). Všechny centrality je možné počítat v orientovaných sítích, ale také u neorientovaných sítí.

Stupňová centralita (Degree Centrality)

Je to základní centralita, která je založena na stupni vrcholu v síti. Určuje se pouze podle tohoto stupně. Pokud je vrchol člověk a spojení mezi dalším člověkem tj. vrcholem je hrana, která určuje to, že se tito lidé znají. Potom je důležitý vrchol (člověk), který zná hodně lidí.

Centralita podle vlastního vektoru (Eigenvector Centrality)

Přirozené rozšíření stupňové centrality, která přiděluje vrcholu jeden bod za každého souseda v síti, nebere se v potaz významnost každého souseda. V centralitě podle vlastního vektoru je centralita vrcholu úměrná součtu centralit jeho sousedů. Podle této metriky je důležitý vrchol, který má hodně sousedů nebo má významné sousedy nebo také obojí.

Centralita vrcholu x_i se počítá jako $x_i = \sum_j A_{ij} x_j$, kde A značí matici sousednosti. Po úpravách x splňuje v limitě rovnici $Ax = K_1 x$, kde K_1 je její největší vlastní číslo (číslo K takové, pro které platí $Ax = Kx$, kde x je vlastní vektor). Centralita pro vrchol i se tedy dá spočítat jako $x_i = K_1^{-1} \sum_j A_{ij}\, x_j$.

Katz- Centralita

Je podobná jako centralita podle vlastního vektoru. Základní myšlenka této centrality je, že každý vrchol zvyšuje centralitu všech sousedních vrcholů o hodnotu úměrnou jeho vlastní centrality. Centralita podle vlastního vektoru naráží na problém v orientovaných sítích, protože vrchol s nulovým vstupním

stupněm má vždy centralitu rovnou nule. Leo Katz přidal do rovnice konstanty, proto i takovéto vrcholy začínají s nenulovou centralitou. Upravená rovnice pro centralitu je $x = \alpha Ax + \beta 1$, kde 1 je vektor samých jedniček. Velice často se počítá $\beta = 1$, protože se nezajímáme přímo o velikost centrality, ale více nás zajímá rozdíl mezi vysokou a nízkou centralitou vrcholu. Rovnice se potom dá přepsat jako $x = (I - \alpha A)^{-1} 1$. Přímý výpočet centralit obnáší invertování matice, proto je tento postup velmi náročný na výpočet, proto se v praxi výpočet provádí iterativně.

PageRank

Tato centralita řeší problém, kdy vrchol s vysokou centralitou sousedící s jinými vrcholy zvyšuje centralitu těchto vrcholů. V centralitě PageRank vrchol zvyšuje centralitu sousedních vrcholů o hodnotu úměrnou jeho centralitě dělenou počtem sousedů tohoto vrcholu. Výsledná rovnice je $x_i = \alpha \sum_j A_{ij} \frac{x_j}{k_j^{out}}$, kde k_j^{out} je výstupní stupeň vrcholu j.

Hubs and Authorities

Výše zmíněné centrality měří důležitost vrcholu v síti podle toho kolik, a jak významných vrcholů na něj ukazuje. Důležitým vrcholem v síti, ale může být i ten který ukazuje na jiné významné vrcholy v síti. Například v síti internetu je to webová stránka, která obsahuje odkazy na stránky k danému tématu, které stránka zkoumá (popisuje). Proto je možné rozlišovat dva druhy důležitých vrcholů a to tzv. *huby* a *autority*. Huby je vrchol, který ukazuje na důležité vrcholy a autorita je vrchol na který ukazuje mnoho vrcholů. Autorita může být hub a také hub může být autoritou. Tuto myšlenku použil Kleinberg, který navrhl algoritmus pro výpočet centralit zvaný *HITS* (Hyperling-Induced Topic Search). Tento algoritmus přiřazuje každému vrcholu Authority centralitu podle toho kolik hub na něj ukazuje a Hub centralitu podle toho, na kolik autorit vrchol ukazuje.

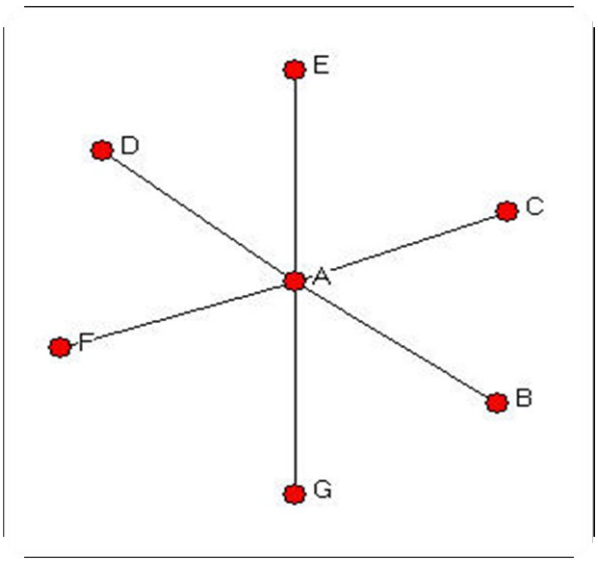
Centralita blízkosti (Closeness Centrality)

Closeness centralita měří významnost vrcholu podle průměrné hodnoty vzdálenosti od všech ostatních vrcholů v síti. Aby významné vrcholy nabyly

vyšší hodnotu, je tato centralita vypočtena jako inverzní hodnota tohoto průměru. Významný vrchol podle této metriky může ostatní vrcholy rychleji ovlivňovat nebo může mít dobrý přístup k informacím o ostatních vrcholech v síti. Průměrnou vzdálenost vrcholu x_i od ostatních vrcholů v síti lze zapsat jako $l_i = \frac{1}{n}\sum_j d_{ij}$, kde n je počet vrcholů v síti a d_{ij} je nejkratší cesta mezi vrcholy x_i a x_j. Výpočet centrality je pak $C_i = \frac{1}{l_i}$.

Centralita mezilehlosti (Betweenness Centrality)

Hodnota centrality pro daný vrchol je počet nejkratších cest mezi každými dvěma vrcholy v grafu sítě, na kterých tento vrchol leží. Pokud v síti dochází ke komunikaci a přes daný vrchol tyto data procházejí, hodnota této centrality vyjadřuje, jaké množství informací přes tento zkoumaný vrchol projde. Vrchol s vysokou centralitou může mít nízký stupeň a také nemusí ležet blízko ostatních vrcholů v síti, podstatné je když přes něj prochází mnoho nejkratších cest v grafu. Tato situace může nastat, když je vrchol tzv. mostem mezi dvěma či více komponentami v grafu sítě. Centralitu mezilehlosti vrcholu x_i lze spočítat jako $B_i = \sum_{st} \frac{n_{st}^i}{g_{st}}$, kde g_{st} je počet všech nejkratších cest mezi vrcholy x_s a x_t, počet nejkratších cest, které vedou přes vrchol x_i je n_{st}^i. [6]



Obrázek 7: Graf ve tvaru hvězdy

(Zdroj: http://www.faculty.ucr.edu/~hanneman/nettext/C10_Centrality.html)

Na obrázku číslo 7 je zachycen graf ve tvaru hvězdy, který znázorňuje extrémní případ Betweenless centrality. Zkoumaný vrchol A je ve středu grafu.

5. Hledání komunit v sociálních sítích

Při hledání komunit v grafu (sociální síti) není předem známo na kolik částí lze graf rozdělit. Velikost jednotlivých částí grafu se může značně odlišovat. A proto je při hledání komunit důležité rozdělit graf na přirozené skupiny vrcholů takovým způsobem, aby mezi těmito skupinami bylo malé množství hran. Vybrané metody hledání komunit jsou popsané níže.

5.1 Metody založené na modularitě

Metody založené na modularitě se nejčastěji používají k rozdělení grafu do dvou komunit. Modularita jako metrika udává rozdíl mezi počtem existujících hran, které jsou mezi vrcholy stejného typu a počtem hran v náhodně vytvořeném grafu v poměru ke všem existujícím hranám v grafu. Vrcholy stejného typu jsou vrcholy, které patří do stejné skupiny (komunity).

Nejjednodušší algoritmus je Kernighan-Lin algoritmus tento algoritmus rozděluje síť do dvou komunit. Algoritmus na začátku náhodně rozdělí graf na dvě stejně velké skupiny. Potom postupně prochází každý vrchol v grafu a vypočítává, jak moc by se změnila modularita, pokud by se daný vrchol přesunul do druhé komunity. Po provedení tohoto kroku algoritmus zvolí vrchol, který nejvíce zvýší nebo sníží modularitu a přesune ho do druhé skupiny. Poté se tento proces opakuje. Je zde ale jedno důležité omezení vrchol, který již byl přesunut, nemůže být přesunut v tomto kole algoritmu. Jakmile jsou všechny vrcholy přesunuty právě jednou, prochází algoritmus zpětně přes stavy, kterými síť prošla a vybere stav s největší modularitou. Tento stav použije algoritmus jako počáteční stav sítě pro opakovaný výpočet. Tento proces se opakuje pořád dokola do té doby, kdy už nedochází ke zlepšení modularity. Algoritmus lze použít po úpravě pro hledání komunit v síti. Úprava je založena na rozdělení každé vzniklé komunity na dvě nové části. Nastává ale zde problém, že ideálního dělení nemusí být dosaženo pomocí nejlepšího rozdělení vždy na dvě části a také je nutné počítat změny modularity pro celou síť a ne jenom modularity v rámci právě dělené komunity.

Běh tohoto algoritmu je reprezentovaný stromem, ve kterém každý vrchol je komunita v určitém kroku algoritmu. Listy jsou jednotlivé vrcholy. Celý stav

sítě odpovídá jednomu patru ve stromě. Jednotlivá patra stromu odpovídají stavu grafu v určitém čase výpočtu. Uživatel se může rozhodnout, který stav použije pro bližší zkoumání. [6]

5.2 Dělení pomocí Betweenness Centrality

Další metodou pro hledání komunit vrcholů v síti je dělení grafu pomocí Betweenness Centrality. Musíme najít hrany, které leží mezi komunitami. Pokud tyto hrany nalezneme a odstraníme je z grafu, tak nám zůstanou pouze izolované komunity, které hledáme. Algoritmus pro detekci komunit pomocí mezilehlosti nejprve spočítá pro všechny hrany v síti betweenness hodnotu. Poté prohledá tyto hrany a hranu s nejvyšší hodnotou vymaže. Vymazání této hrany změní hodnotu centrality některých hran. Proto musí následovat po odstranění hrany s největší hodnotou centrality nový výpočet centralit mezilehlosti. Po tomto kroku se algoritmus opakuje. Během tohoto procesu se bude síť postupně „rozpadat“ na více částí. Tímto postupem získáme hledané komunity. Tato metoda je pomalejší než metody založené na modularitě, poskytuje však srovnatelně kvalitní výsledky. Průběh tohoto algoritmu je možné reprezentovat stromem, jak již bylo zmíněno dříve. [6]

5.3 Dělení podle metod hierarchického shlukování

Hierarchické shlukování (Hierarchical Clustering) je třída metod, ve které se používají algoritmy, které pracují s modularitou. Hierarchické shlukování je aglomerativní technika, ve které se jednotlivé vrcholy sítě spojují dohromady, aby vytvořili skupiny. Základní myšlenkou je definovat míru podobnosti nebo sílu propojení mezi vrcholy, založené na struktuře sítě a poté spojit dohromady nejbližší či nejvíc podobné vrcholy do skupin. Pro rozhodnutí, které vrcholy spojit se používá metrika míry podobnosti vrcholů. Míra podobnosti porovnává pouze dvojice vrcholů.

Míru podobnosti skupin lze vytvořit třemi způsoby:

- single-linkage – podobnost dvou skupin je definována jako podobnost dvojice jejich nejpodobnějších vrcholů

- complete-linkage – podobnost skupin je rovno podobnosti dvojice nejméně podobných vrcholů
- average-linkage – podobnost skupin je rovno průměrné podobnosti všech dvojic vrcholů

Average – linkage (průměrné propojení) je nejvíce používané, jelikož určitým způsobem zahrnuje celkovou podobnost obou skupin a ne pouze extrémy

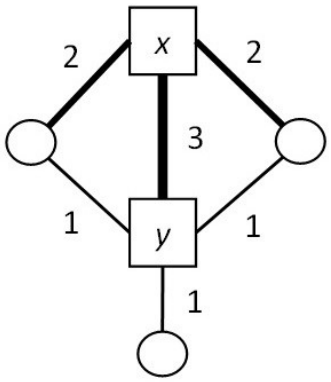
Výstupem tohoto postupu je opět strom. Lze takto najít subkomunity uvnitř komunit a jiné cenné informace o základní struktuře zkoumané sítě.[6]

5.4 Hledání komunit pomocí lokální expanze

Jelikož jsem si tuto metodu pro hledání komunit v síti vybral pro implementaci v mé aplikaci. Budu tento postup hledání komunit popisovat vcelku podrobněji níže. Jsou zde popsány jednotlivé kroky, které se musí aplikovat jako základ pro fungování algoritmu.

5.4.1 Měření závislosti mezi dvěma vrcholy

Pojem závislosti mezi dvěma vrcholy si lze vysvětlit na obrázku níže. Jsou zde dva sousední vrcholy *x* a *y*, které jsou znázorněny ve váženém neorientovaném grafu. Síla spojení mezi dvěma vrcholy je znázorněna hodnotou váhy jejich hrany. Když budeme uvažovat o závislosti, tak můžeme předpokládat, že závislost mezi vrcholem *x* a *y* z části ovlivní i vrcholy přes které jsou nepřímo propojeny. Proto je třeba uvažovat jak tuto závislost změřit. [25]



Obrázek 8: Příklad závislosti mezi dvěma vrcholy

Řekněme, že $E(x)$ je množina všech hran přilehlých k vrcholu *x*. Potom $Adj(x,y)$ je množina všech hran mezi vrcholem *x* a kterýmkoliv vrcholem sousedícím s vrcholem *y*. Přesněji $Adj(x,y) \subseteq E(x)$. $W(e)$ je váha hrany *e* a $W(v_1,v_2)$ je váha hrany mezi vrcholem v_1 a vrcholem v_2. Pokud je váha mezi vrcholy v_1 a v_2 rovna nule tak to znamená, že mezi nimi neexistuje žádná hrana. Pokud *x* není izolovaný vrchol v síti. Pak je závislost $D(x,y)$ vrcholu *x* na vrchol *y* definována takto:

$$D(x,y) = \frac{W(x,y) + \sum_{e_i \in Adj(x,y)} W(e_i) \cdot R(e_i)}{\sum_{e_i \in E(x)} W(e_i)},$$

$$R(e_i) = \frac{W(y,v_i)}{W(e_i) + W(y,v_i)}.$$

$R(e_i)$ je koeficient závislosti vrcholu x na vrcholu y přes společné sousední vrcholy v_i a proto z toho plyne $v_i \in e_i$.

Výpočet pro situaci na obrázku číslo 8.

$$D(x,y) = \frac{3 + 2 \cdot \frac{1}{2+1} + 2 \cdot \frac{1}{2+1}}{2+3+2} = \frac{13}{21}$$

Tato závislost popisuje vztah vrcholu x vůči vrcholu y a jeho okolí. V tomto případě vypočítaná závislost je větší než 0,5. Z použité rovnice lze odvodit $D(x,y) \in \langle 0;1 \rangle$. Pokud výsledek závislosti nabývá hodnoty rovnající se nule, znamená to, že vrchol *x* a vrchol *y* nemají společnou hranu nebo společný sousední vrchol tj. nejsou spolu vůbec propojeni. Výsledek závislosti roven 1 popisuje situaci, ve které je vrchol *x* spojen pouze jednou hranou s vrcholem *y*.

5.4.2 Měření závislosti vrcholu na množině vrcholů v síti

Vztah závislosti vrcholu na jiném může být popsán jako závislost vrcholu *x* na množinu *n* vrcholů z $Y = \{y_1, y_2, \ldots, y_n\}$. V tomto případě musíme vzít v potaz dvě omezení. Zaprvé vrchol *x* může mít více sousedů z množiny *Y*, a proto množiny těchto sousedů jsou $N(x,Y)$, $N(x,Y) \subseteq E(x)$. Zadruhé $Adj(x,Y)$ je množina všech hran obsahujících vrchol x a jednoho ze sousedů v_i, $i = 1,\ldots,n$, z vrcholu z množiny *Y*, potom $v_i \notin N(x,Y)$. Jakýkoliv soused v_i vrcholu *x* může být sousedem více než jednoho vrcholu z y_j, $j = 1,\ldots,n$, $i \neq j$ z množiny Y ve stejnou dobu. Tato množina vrcholu je $Y(v_i)$. Vezměme případ, že *x* není izolovaný vrchol sítě. Potom závislost $D(x,Y)$ vrcholu *x* na množinu vrcholů *Y* je definována takto:

$$D(x,Y) = \frac{\sum_{y_i \in N(x,Y)} W(x,y_i) + \sum_{e_i \in Adj(x,Y)} W(e_i) \cdot R(e_i)}{\sum_{e_i \in E(x)} W(e_i)},$$

$$R(e_i) = MAX_{yj \in Y(v_i)} \left(\frac{W(y_j, v_i)}{W(e_i) + W(y_j, v_i)} \right).$$

Hodnota koeficientu závislosti $R(e_i)$ je maximum z hodnot vypočítaných pro jednotlivé vrcholy z množiny $Y(v_i)$, které mají vzájemný vrchol v_i s hranou e_i jejichž druhý vrchol je vrchol x.

5.4.3 Detekce komunity založená na vrcholové závislosti

Před popisem samotného algoritmu je potřeba vysvětlit použité značení a pojmy.

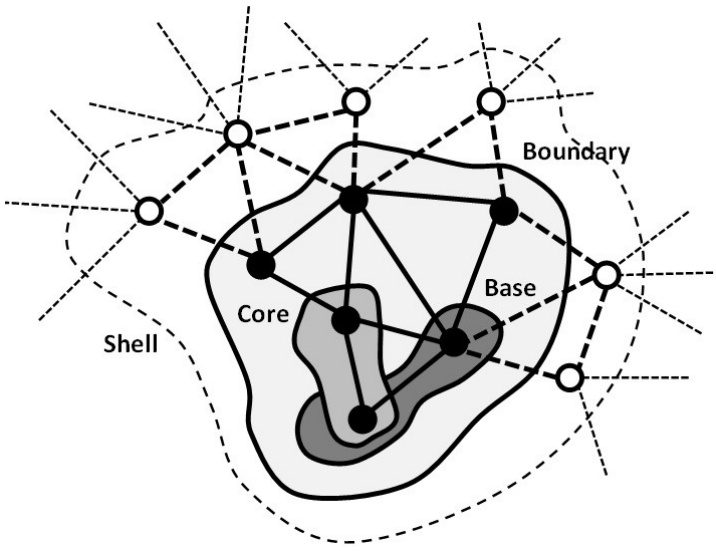
Community base – je to počáteční sada n vrcholů vhodně vybraných předem, které podle definice patří ke komunitě a splňuje dvě kritéria. Zaprvé tato sada vrcholů tvoří bikonektní podgraf a zadruhé, alespoň $(n-1)$ vrcholů je závislých na jiné základní vrcholy.

Community boundary B – hranice komunity, je tvořena všemi vrcholy, které patří do komunity L, ale nepatří do community core C. Protože alespoň jedna hrana je orientovaná ven z komunity L.

Community core C – jádro komunity, zahrnuje všechny vrcholy, které nemají žádnou hranu orientovanou ven z komunity L (ven z C a B).

Community shell S – „skořápka“ komunity, zahrnuje nepoznané síťové vrcholy, které se stanou poznanými a mohou být přesunuty do komunity L v průběhu procesu lokální expanze.

Community L – komunita, je tvořena všemi vrcholy z C a B tj. $L = C \cup B$



Obrázek 9: Lokální komunita

Algoritmus je založen na postupném přezkoumávání vrcholů z shell community, následného přiřazení vrcholů, které splňují kritéria pro přiřazení

do komunity. Pouze vrcholy, které nesplňují kritéria pro přiřazení zůstávají v shell S. Vrcholy, které splňují kritéria se přiřadí do community boundary B. Následně pak vrcholy, které patří do nového shell, jsou přidány a proces celého výpočtu se znovu opakuje.

Popis jednotlivých kroků, které algoritmus provádí:

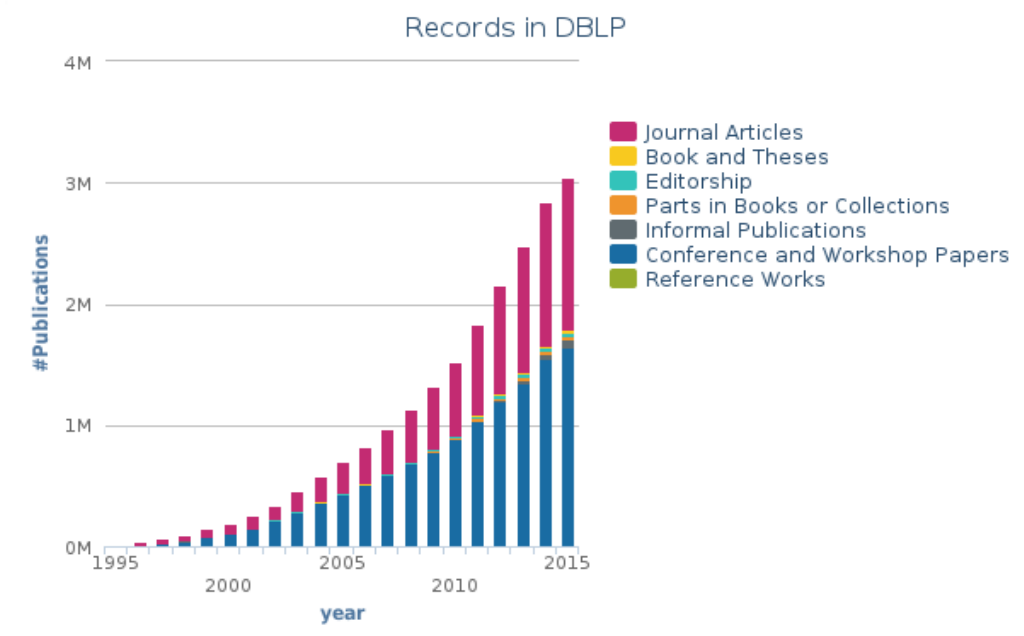
1. Vytvoří community boundary B, které je rovno community base.
2. Vytvoří prázdné community core C.
3. Vytvoří prázdné community shell S.
4. Přesune vrcholy z community boundary B, které nemají sousedy vně community L, do community core C.
5. Znovu naplní community shell novými sousedy vrcholů přidaných do community boundary B, které jsou vně community L.
6. Spočítá závislosti na vrcholy komunity pro každý vrchol z community shell S.
7. Přesune každý vrchol z community shell S, který splňuje kritéria pro přestěhování a uznání, že patří do komunity L do community boundary B.
8. Pokud alespoň jeden vrchol z community shell S je rozpoznán (uznán tj. přesunut do community L), opakuje výpočet algoritmu od bodu 4.
9. Community L je sjednocení community core a community boundary a proto platí $L = C \cup B$.[25]

6. Praktická část práce

V praktické části pracuji s mnou navrženou aplikaci, která používá pro hledání komunit algoritmus lokální expanze, který je podrobně popsán v teoretické části této diplomové práce. Pomocí této aplikace jsem provedl experiment na testovacích datech z databáze DBLP.

6.1 Databáze DBLP

Pro testování aplikace jsem vybral databázi DBLP. Je to databáze publikací z velkých odborných konferencí a odborných časopisů. DBLP databáze začala vznikat v roce 1993 jako malý experimentální Web Server, který se v toku času rozvinul do služby, která je značně užívaná vědeckou komunitou. Pro názornost jak roste využití databáze v toku času lze vidět na obrázku níže. Tato databáze, je používána jako testovací databáze pro nové algoritmy, kdy není důležitý její obsah, ale je možné z ní sestrojit různé druhy grafů, například graf spoluautorství, který tvoří sociální síť. Tato síť spoluautorství je vážená síť, kde váha vztahu je dána počtem společných publikací obou autorů. Tyto vztahy jsou oboustranné, a proto je síť neorientovaná. [23]



Obrázek 10: Graf publikací v databázi DBLP podle roků

(Zdroj: http://dblp.uni-trier.de/statistics/recordsindblp.html)

6.2 Aplikace

6.2.1 Specifikace aplikace

Aplikace má sloužit ke zkoumání testovací databáze spoluautorské sítě DBLP. Musí umět vypočítat významnost vrcholů z této testovací databáze a následně vyhledat komunity. Výsledky těchto výpočtů musí přehledně zobrazit a také nabídnout možnost pozorování vývoje komunit kolem konkrétního autora v toku času.

6.2.2 Funkční požadavky

- Načtení databáze ze souboru
- Výpočet všech komunit z databáze
- Zobrazení výsledků výpočtů
- Zobrazení komunit(y), do kterých patří zvolený vrchol (autor)
- Uložení zobrazeného grafu komunity do souboru
- Uložení nalezených komunit do souboru

6.2.3 Ostatní požadavky

- .Net Framework 4.5.1
- MS Windows Vista SP2 a vyšší verze
- Procesor 1GHz
- Paměť RAM minimálně 512MB

6.2.4 Použité knihovny v programu

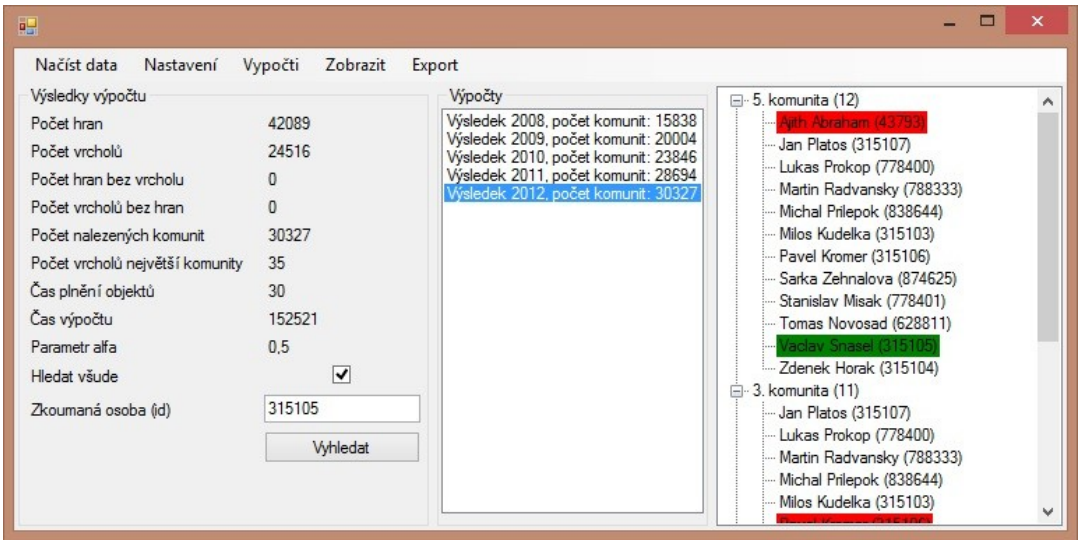
Pro vykreslování grafů v programu byly použity knihovny Microsoft.MSAGL.dll, Microsoft.MSAGL.Drawing.dll a Microsoft.MSAGL.GraphViewerGDIGraph.dll. Tyto knihovny slouží přímo k vykreslení grafu, poskytují nastavení atributů vykreslení a to například nastavení barvy čar, styly čar, nastavení tvarů a barev uzlů.

```
using Dip_console;
using System;
using System.Collections.Generic;
using System.Windows.Forms;
1 reference
class GraphicGraph
{
    1 reference
    public static void CreateGraph(List<Edge> edges)
    {
        //create a form
        System.Windows.Forms.Form form = new System.Windows.Forms.Form();
        //create a viewer object
        Microsoft.Msagl.GraphViewerGdi.GViewer viewer = new Microsoft.Msagl.GraphViewerGdi.GViewer();
        //create a graph object
        Microsoft.Msagl.Drawing.Graph graph = new Microsoft.Msagl.Drawing.Graph("graph");
        //create the graph content
        foreach (var item in edges)
        {
            graph.AddEdge(item.Vertices[0].Name + " (" + item.Vertices[0].Id + ")", item.Vertices[1].Name +
        }


        viewer.Graph = graph;
        //associate the viewer with the form
        form.SuspendLayout();
        viewer.Dock = System.Windows.Forms.DockStyle.Fill;

        viewer.LayoutEditingEnabled = false;
        form.Controls.Add(viewer);
        form.ResumeLayout();
        //show the form
        form.ShowDialog();
    }
}
```

Obrázek 11: Ukázka kódu programu ve statické třídě

6.2.5 Implementace

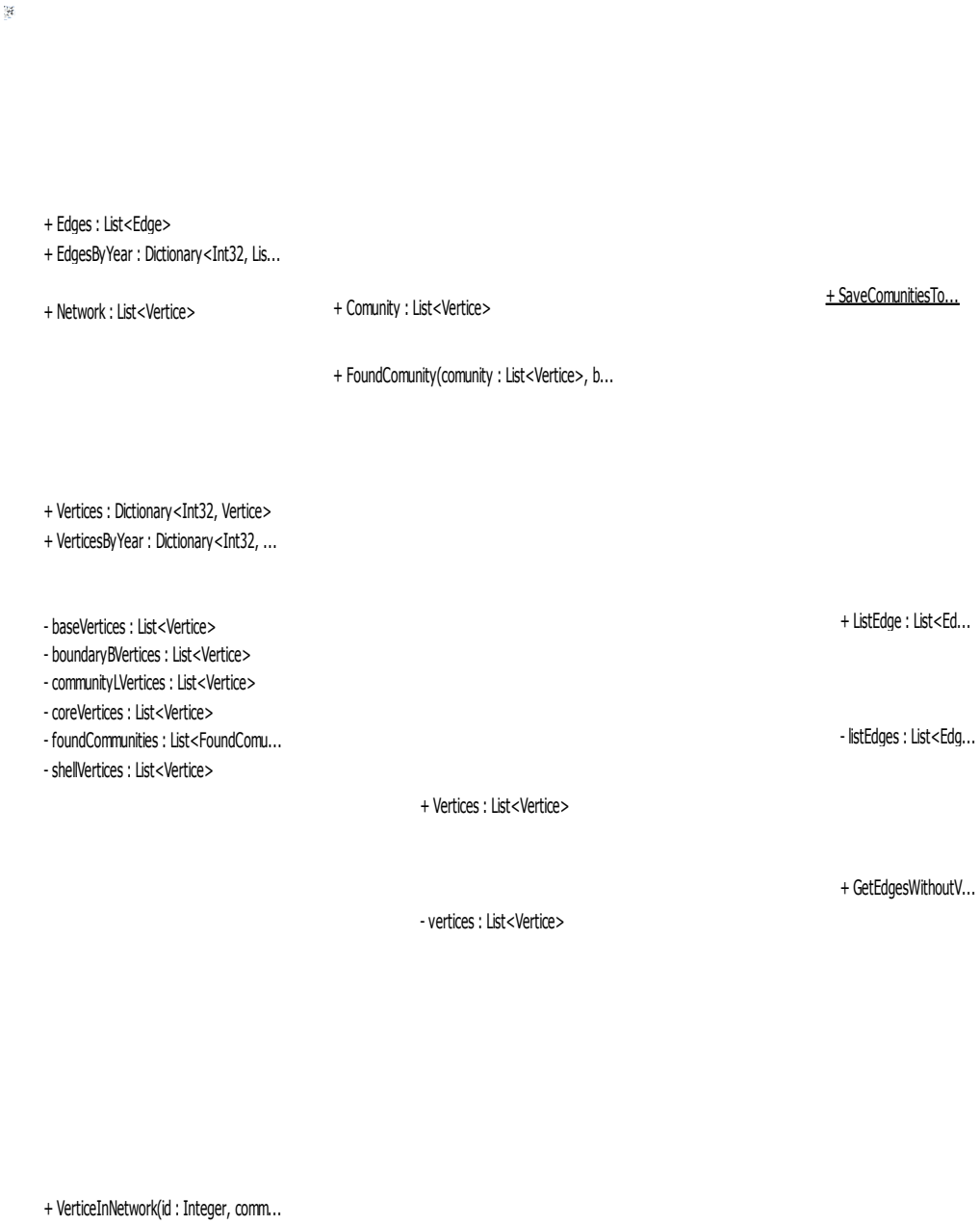
Aplikaci jsem napsal v jazyce C#, grafické rozhraní je implementováno ve Windows Forms, což je grafická nadstavba, která pouze obaluje Windows API. Windows Forms je vyladěná a osvědčená nadstavba, která v základu obsahuje množství grafických prvků. Jako vývojové prostředí bylo použito Visual Studio 2013. Aplikace je 32 bitová a je rozdělená na dvě vrstvy a to na logiku programu a grafické rozhraní.



Obrázek 12: Hlavní okno aplikace

6.2.6 Diagram hlavních tříd programu

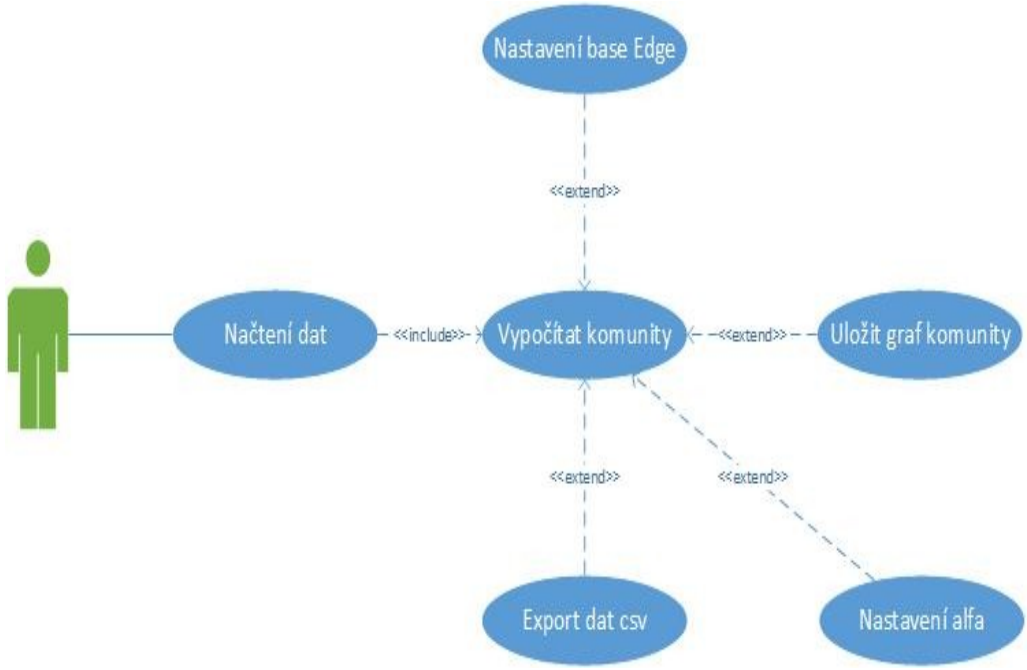
Diagram popisuje hlavní logiku programu. Objekt Graf obsahuje kolekci všech vrcholů a hran. V tomto objektu jsou také implementované metody pro vypočtení algoritmů. Objekt Edge obsahuje přesně dva vrcholy a objekt Vertice obsahuje x hran. Hlavní logika programu je v těchto třech objektech.



Obrázek 13: Diagram tříd

6.2.7 Use case diagram aplikace

Diagram případů užití zobrazuje chování programu tak, jak ho vidí uživatel, v tomto případě zelená postavička. Účelem diagramu je popsat funkce programu, tedy to co od něj uživatel očekává. Diagram zobrazuje co má systém umět, ale nevypovídá nic o tom, jak to dělá.



Obrázek 14: Use case diagram

6.3 Experiment

Experiment a jeho výsledky jsou založeny na analýze dat, které tvoří největší souvislá komponenta vážené spoluautorské sítě založená na datech z databáze DBLP od roku 2008 do roku 2012, která je filtrovaná tak, aby v ní nebyli nevýznamní autoři. To znamená, že je síť zbavená tzv. šumu. Na základě získaných informací jsou vytvořené jednotlivé statistiky vývoje komunit podle roků.

6.3.1 Testovací hardware

Experiment byl prováděn na následující konfiguraci:

- 2 jádrový procesor Intel Core i5-4200U 1.6GHz
- 4 GB operační paměti
- Windows 8.1 Professional 64bit

6.3.2 Statistiky

Rok	Hrany	Vrcholy	Nalezené komunity	Největší komunita	Čas výpočtu
2008	21741	13247	15838	29	35,06
2009	27105	16534	20004	37	53,4
2010	33243	19664	23846	47	105,09
2011	39635	23510	28694	40	108,01
2012	42089	24516	30327	35	144,82

Tabulka 1: Výpočet vše komunit podle roků

Uvedená tabulka popisuje počet vstupních hran, vrcholů a počet nalezených komunit dle roků. Také obsahuje údaj o počtu autorů v největší komunitě a celkový čas výpočtu všech nalezených komunit dle roků v sekundách.

6.3.3 Vývoj komunity kolem vybraného autora v čase

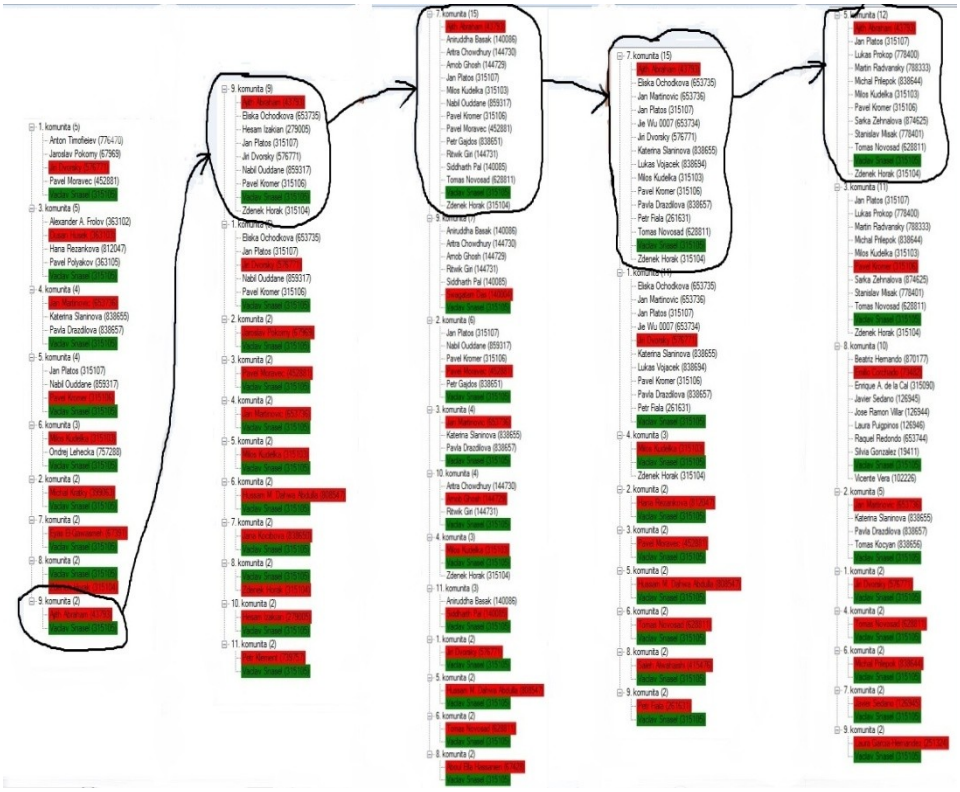
Konkrétní zkoumaný významný autor je prof. RNDr. Václav Snášel, CSc. Je od roku 2009 děkanem Fakulty elektrotechniky a informatiky až doposud. Publikoval více, než 400 prací z toho 201 prací je zaznamenáno na Web of Science a 268 je zaznamenáno ve SCOPUS. Také má 1482 citací na Google Scholar. [27]

Zkoumaný autor v období od roku 2008 do roku 2012 vytvořil mnoho komunit s ostatními autory. V tabulce níže lze tento fakt zpozorovat. Tabulka zobrazuje jednotlivé zkoumané roky.

Zkoumaný rok	Počet všech komunit	Největší komunita
2008	9	5
2009	11	9
2010	11	15
2011	9	15
2012	9	12

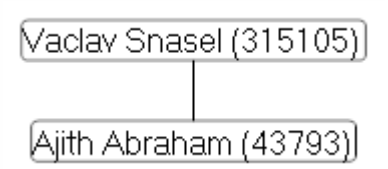
Tabulka 2: Komunity kolem zkoumaného autora

Z tabulky číslo 2 je zřejmé, že v roce 2008 zkoumaný autor tvořil největší komunitu o 5 členech, ale tato komunita se následující rok rozpadla. Proto jsem si pro popis vybral vznikající komunitu s autorem Ajith Abraham (43793).



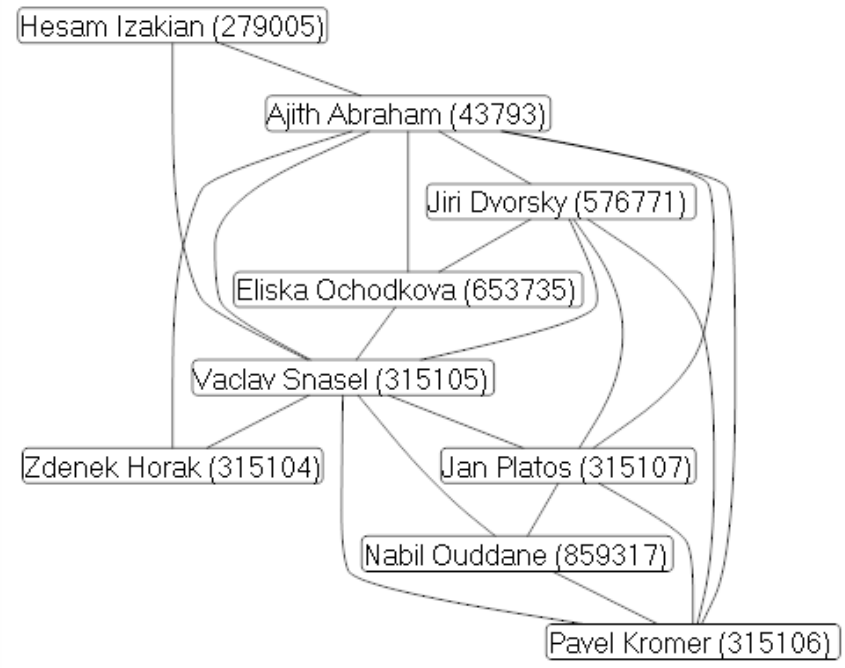
Obrázek 15: Vybraná komunita

V roce 2008 začal spolupracovat Václav Snášel (315105) s Ajith Abraham (43793). Vytvořili spolu komunitu, která má pouze dva vrcholy a jednu hranu. V tomto případě tito dva autoři tvoří komunitu L a současně také tvoří community base.



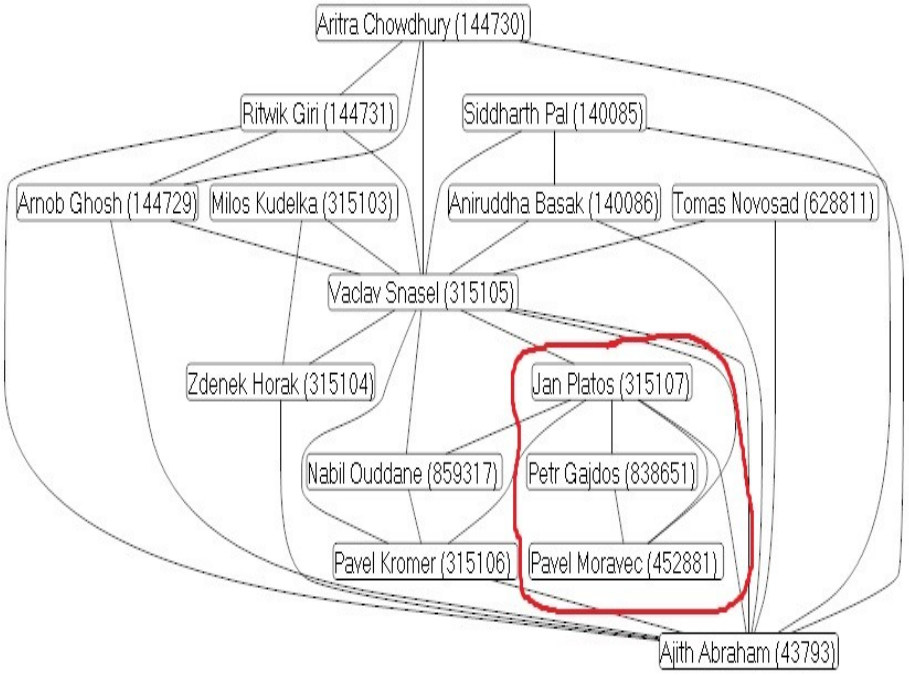
Obrázek 16: Komunita kolem V. Snášel 2008

V roce 2009 se komunita rozrostla o 7 dalších spoluautorů. Z grafu lze vyčíst, že autor Václav Snášel spolupracoval se všemi autory přímo.



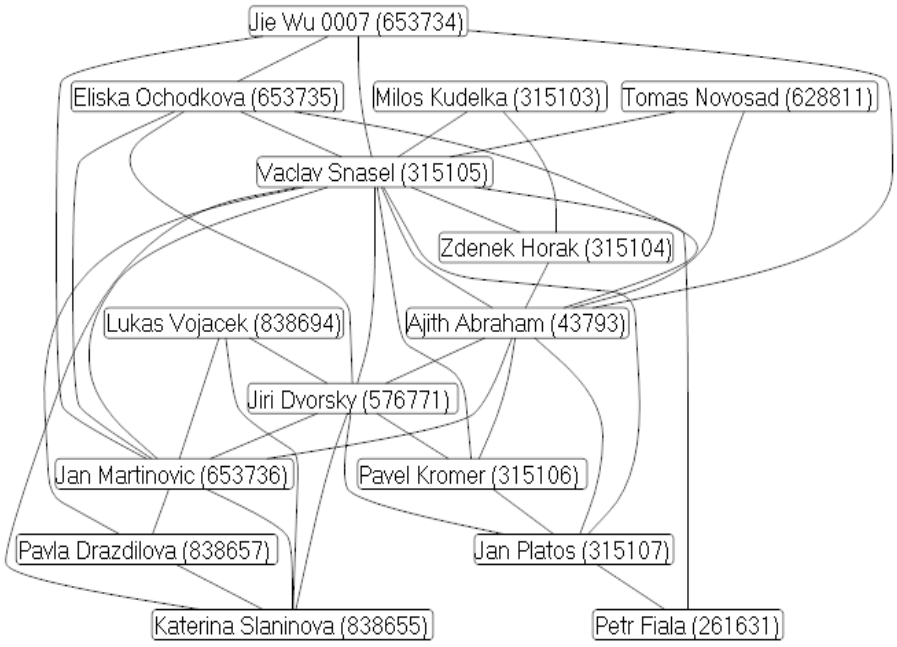
Obrázek 17: Komunita kolem V. Snášel 2009

V roce 2010 komunita kolem zkoumaného autora měla celkově 15 členů. Z grafu je zřejmé, že Petr Gajdoš není přímo spojen se zkoumaným autorem Václavem Snášelem. I přesto, je členem této komunity, jelikož patří do takzvané subkomunity, která má 3 členy viz. obrázek níže.



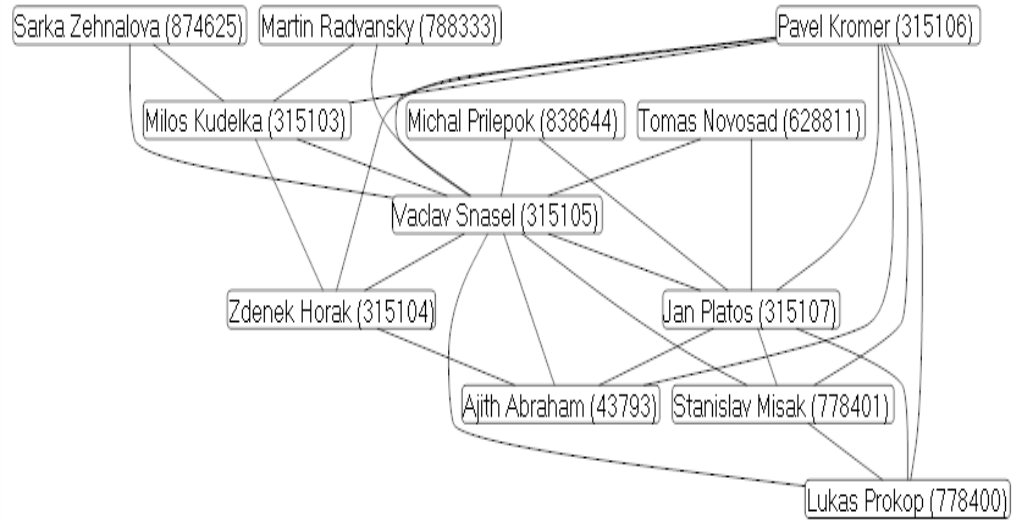
Obrázek 18: Subkomunita kolem V. Snášel 2010

Rok 2011 komunita má stále 15 členů, ale někteří autoři komunitu opustili (Aniruddha Basak, Aritra Chowdhury, Arnob Ghosh, Nabil Ouddane, Pavel Moravec, Petr Gajdoš, Ritwik Giri, Siddharth Pal) zřejmě ukončili spolupráci na odborném textu.



Obrázek 19: Komunita kolem V. Snášel 2011

Rok 2012 komunita kolem zkoumaného autora má celkem 12 členů.



Obrázek 20: Komunita kolem V. Snášel 2012

7. Závěr

V jednotlivých kapitolách práce jsem se zaměřil na popis sociálních sítí, jejích historický vývoj a uvedl jsem příklady nejznámějších sociálních sítí. Popsal jsem vybrané metody hledání komunit v grafu sítě. Také jsem vytvořil aplikaci, ve které je implementován algoritmus, který používá pro hledání komunit v síti metodu lokální expanze. Pomocí této aplikace byl proveden výpočet nad testovací databází. Získané výsledky jsem uvedl ve formě tabulky, ze které lze vyčíst, jak rychle algoritmus pracuje. Provedl jsem analýzu vývoje komunity kolem vybraného autora, v období od roku 2008 do roku 2012.

Seznam zdrojů

[1] SOLOMON. G., SCHRUM. L., *WEB 2.0 new tools, new schools.* First edition. WASHINGTON D.C. Book publishing. 2007. ISBN: 978-1-56484-234-3.

[2] Srov. BAACK, D., CLOW, K.E. *Reklama, propagace a marketingová komunikace*. 1.vyd. Praha: Computer Press, 2008. ISBN 978-80-251-1769-9.

[3] BOUČKOVÁ, J. a kol. *Marketing.* 1.vyd. Praha: C.H.Beck, 2003. ISBN 80-7179-577-1.

[4] BOUČKOVÁ, J. a kol. *Základy marketingu*. 4. vyd. Praha: Oeconomica, 2011. ISBN 8024517604.

[5] PŘIKRYLOVÁ, J. *Moderní marketingová komunikace.* Praha: Grada Publishing, a.s., 2010., ISBN 978-80-247-3622-8.

[6] NEWMAN, M.E.J. *Networks an introduction.* New York: Published in the United States by Oxford University Press Inc., 2010. ISBN 978-0-19-920665-0.

[7] Aktuálně. *Sociální sítě.* [online]. ©1999 - 2015 [cit. 2015-06-01]. Dostupné z WWW: http://www.aktualne.cz/wiki/veda-a-technika/socialni-site/r~i:wiki:1456.

[8] LIBTON C. FREEMAN, *Miles.Visualizing Sosial Networks.* [online]. ©2011 [cit. 2015-10-02]. Dostupné z WWW: http://www.cmu.edu/joss/content/articles/volume1/Freeman.html.

[9] Ihned. *Pan Facebook: Soukromí je přežitek 20.století.* [online]. ©2010 [cit. 2014-12-11]. Dostupné z WWW: http://archiv.ihned.cz/c1-43424300-pan-facebook-soukromi-je-prezitek-20-stoleti.

[10] app4page. *Počet uživatelů na Facebooku – Demografické rozdělení.* [online]. ©2014 [cit. 2014-12-11]. Dostupné z WWW: http://app4page.com/cz/blog/pocet-uzivatelu-facebook-demograficke-rozdeleni-cr/.

[11] HANDL. J., *Linkedln pro začátečníky.* [online]. ©2009 [cit. 2014-12-11]. Dostupné z WWW: http://www.lupa.cz/clanky/linkedin-pro-zacatecniky.

[12] Business2community. *Jak používat Twitter pro podnikání.* [online]. ©2014 [cit. 2014-12-11]. Dostupné z WWW: http://www.business2community.com/twitter/use-twitter-business-2-01077755.

[13] Otevřená encyklopedie. *My Space*. [online]. ©2015 [cit. 2015-06-06]. Dostupné z WWW: http://en.wikipedia.org/wiki/Myspace.

[14] DOČEKAL. D., *Infografika: Světová mapa sociálních sítí z roku 2011*. [online]. ©2013 [cit. 2014-20-11]. Dostupné z WWW: http://www.justit.cz/wordpress/2013/05/29/infografika-svetova-mapa-socialnich-siti-z-roku-2011/.

[15] WALKER, *Miles. The History of Social Networking*. [online]. ©2009-2015 [cit. 2015-12-02]. Dostupné z WWW: http://www.webmasterview.com/2011/08/social-networking-history/.

[16] Interent Raly Chat, *Interent Raly Chat.* [online]. ©2013 [cit. 2015-01-02]. Dostupné z WWW: http://cs.wikipedia.org/wiki/Internet_Relay_Chat.

[17] BERNERS-LEE, *Sir Timothi. Longer Biography* [online] ©2015 [cit. 2015-12-01]. Dostupné z WWW: http//www.w3.org/people/Berners-Lee/Longer.html.

[18] Redakční systémy, *Webovky, blogy, eshopy, for a – CMS redakční a publikační systémy.* [online]. ©2015 [cit. 2015-13-01]. Dostupné z WWW: http://phprs.cz/historie-sluzeb-pro-tvorbu-stranek/.

[19] Probyznys. *Sociální sítě a jejich firemní využití.* [online]. ©2012 [cit. 2014-28-11]. Dostupné z WWW: http://probyznysinfo.ihned.cz/c1-57490210-socialni-site-a-jejich-firemni-vyuziti-jake-jsou-jejich-vyhody-a-nevyhody.

[20] Článek bezkškálová síť. *Bezkškálová síť.* [online]. ©2015 [cit. 2015-22-03]. Dostupné z WWW: https://cs.wikipedia.org/wiki/Bezškálová_síť.

[21] NEWMAN. M., E., J.. *The structure and function of complex networks*, SIAM Review, SIAM, [online]. ©2004 [cit. 2014-22-11]. Dostupné z WWW: http://www-personal.umich.edu/~mejn/courses/2004/cscs535/review.pdf.

[22] Onas seznam. *Spolužáci.* [online]. ©1996 - 2015 [cit. 2014-11-12]. Dostupné z WWW http://onas.seznam.cz/cz/spoluzaci-cz.html.

[23] Dblp. *Computer science bibliografy.* [online]. ©2015 [cit. 2015-05-01]. Dostupné z WWW: http://dblp.uni-trier.de.

[24] Otevřená encyklopedie .*SixDegrees.* [online]. ©2015 [cit. 2015-06-01]. Dostupné z WWW: http://en.wikipedia.org/wiki/SixDegrees.com.

[25] Odborný článek VŠB. *Local Community Detection and Visualization: Experiment Based on Student Data*. Autoři: Miloš Kudělka, Pavla Dráždilová, Eliška Ochodková, Kateřina Slaninová, Zdeněk Horák

[26] Cvrček. M., *Web 2.0* [online]. ©2008 [cit. 2014-15-11]. Dostupné z WWW: http://www.onlio.com/clanky/web_2.0.html

[27] VŠB-TUO, *prof. RNDr. Václav Snášel, CSc.,* [online]. Dostupné z WWW: http://snasel.vsb.cz/en/index.php

Seznam obrázků